DOCUMENT RESUME

ED 466 776 TM 034 287

AUTHOR	Ray, Janet
TITLE	Meta-Analytic Thinking: Using ESCI To Synthesize Effects across Studies.
PUB DATE	2002-02-16
NOTE	32p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, February 14-16, 2002).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Effect Size; *Meta Analysis; Statistical Significance; *Synthesis
IDENTIFIERS	*Confidence Intervals (Statistics)

ABSTRACT

The practical significance, usefulness, and generalizability of research have for years hinged on a finding of statistical significance. Voices of reform have called for the use of effect sizes, confidence intervals, and meta-analytic synthesis of research as a way to judge the practical significance and generalizability of a discovery. This paper discusses the role of meta-analysis in the development of knowledge and the additional power of meta-analytically examining confidence intervals for effects. The Exploratory Software for Confidence Intervals (G. Cuming and S. Finch, 2001; ESCI) software is introduced as a means of synthesizing confidence intervals for effects from past research literature, the effect confidence interval for a current study, and the final pooling of a present study with prior research. (Contains 1 table, 2 figures, and 36 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 466 776

Running head: META-ANALYTIC THINKING

Meta-analytic Thinking: Using ESCI to Synthesize Effects Across Studies

Janet Ray

University of North Texas

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. Ray
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

TM034287

Paper presented at the annual meeting of the Southwest Educational Research Association, February 16, 2002, Austin.

BEST COPY AVAILABLE

Abstract

The practical significance, usefulness, and generalizability of research has for decades hinged on a finding of statistical significance. Voices of reform have called for the use of effect sizes, confidence intervals, and meta-analytic synthesis of research as a way to judge the practical significance and generalizability of a discovery. This paper discusses the role of meta-analysis in the development of knowledge and the additional power of meta-analytically examining confidence intervals for effects. The Exploratory Software for Confidence Intervals (ESCI) software is introduced as a means of synthesizing confidence intervals for effects from past research literature, the effect confidence interval for a present study, and the final pooling of a present study with prior research.

Meta-analytic Thinking: Using ESCI to Synthesize Effects Across Studies

Discovery drives scientific inquiry. We want it to make a difference in how our field of interest operates. So how do we determine the import of a scientific discovery? How do we know if our discovery has meaning outside our own research setting? Peruse the major educational and psychological journals of the past six or seven decades, and the equivalent of importance is usually *statistically significant* (Schmidt, 1992).

Statistical significance testing (SST), sometimes written as null hypothesis statistical testing (NHST), presents as an objective, scientific means of handling accumulated knowledge (Cortina & Dunlap, 1997; Hubbard & Ryan, 2000; Kirk, 1996; Kover, 2000; Stewart, 2000). For some time, however, many authors have refuted prevalent interpretations of SST and promoted increased focus on practical significance and generalizability (cf., Cohen, 1994; Schmidt, 1996; Thompson, 1994,1996). Momentum for reforming the use of SST has grown in recent years, as well as for a more hands-on, reflective, and theory-driven interpretation of data. The inclusion of effect sizes and confidence intervals in research reporting has been encouraged as a way to more fully judge the practical significance of a discovery. In addition, meta-analytic synthesis of research findings has been advocated as a way to build a knowledge base, construct theory, and assess the generalizability of findings.

The purpose of the present paper is to discuss the role of meta-analysis in the development of knowledge and the additional power of meta-analytically examining confidence intervals for effects. In addition, the utility of a new software program, Exploratory Software for Confidence Intervals (ESCI) (Cumming & Finch, 2001), in the synthesis of confidence intervals from past research literature, a present study, and a

pooling of past and present studies will be demonstrated. This paper will (a) discuss the importance of reporting and interpreting effect sizes and confidence intervals in determining the practical significance of research, (b) examine the role of meta-analysis in research synthesis, (c) present meta-analytic thinking as a superior means of interpreting research results, and (d) use the ESCI software to demonstrate a synthesis of confidence intervals from past and present studies.

Effect Sizes, Confidence Intervals, and Practical Significance

Kirk (2001) summarized the researcher's objectives with three questions: "(a) Is an observed effect real or should it be attributed to chance? (b) If the effect is real, how large is it? and (c) Is the effect large enough to be useful?" (p. 213). Traditionally, SST has been used by some to provide answers to all three of these questions. However, Cohen (1994) emphatically contended that SST is not the final answer and noted that null hypothesis significance testing "does not tell us what we want to know" (p. 997). What, then, can answer questions about chance, magnitude of effect, and usefulness?

Statistical Significance versus Practical Significance

It is beyond the scope of this paper to exhaustively examine the advantages and disadvantages of SST or its standing in historical social science research. However, in order to understand what meta-analytic thinking can do, we need to understand what SST can and cannot do, specifically in the context of the practical significance of research findings.

Thompson (1994) explained that statistical significance answers this question: "Assuming the sample data came from a population in which the null hypothesis is (exactly) true, and given our sample size(s), is the calculated probability of our sample

results less than the acceptable limit (p(CRITICAL)) imposed regarding Type I error?" (p. 2). Note that in the definition of SST, the null hypothesis is always assumed to be exactly true because the population is unknown. In actuality, the null is never exactly true in the population as there will always be differences/relationships in the population, no matter how trivial (Thompson, 1996). If enough subjects are collected and the researcher has a sufficiently strong statistical microscope, the null will always be found false (Cohen, 1994; Henson & Smith, 2000; Thompson, 1994, 1998). Kirk (1996) illustrated this principle in a scenario of two researchers who employ identical treatments, but disagree on the statistical significance of their studies. One researcher does not reject the null at .06, but the other researcher, with a slightly larger sample, decides to reject the null at .05. Kirk exclaimed with Rosnow and Rosenthal (1989), "Surely God loves the .06 as much as the .05" (p. 1277). Statistical significance tests, therefore, only partially answer the first question posed by Kirk (2001): the probability (chance) of getting the present results for the *present sample*, if we assume the null is exactly true in the population.

Knowledge about the magnitude of an observed effect is crucial in determining the usefulness of research results. A small p value (the calculated probability of an observed result) has been widely misinterpreted as an inverse indicator of effect size (Finch, Cumming, & Thomason, 2001). Kirk (1996) observed:

> Because the null hypothesis is always false, a decision to reject it simply indicates that the research design had adequate power to detect a true state of affairs, which may or may not be a large effect or even a useful effect. (p. 747)

In fact, *p* values tell us nothing about the magnitude of an observed effect. As noted by Bailar and Mosteller (1988), "Merely reporting a *P* value from a significance test of differences loses the information about both the average level of performance and the variability of individual outcomes for the separate treatments" (p. 267).

Usefulness of findings underlies Kirk's (2001) three objectives for the researcher. If the researcher cannot make a case for the presence of the observed effect in the population of interest, the usefulness of the findings are suspect. Herein lies the problem with relying on SST as the standard of usefulness: SST evaluates the likelihood of the effect in the sample, not the population (Vacha-Haase, 2001). In SST, we assume the null hypothesis to be true in the population and then determine the probability of our sample results. Replicability, and thus usefulness, cannot be evaluated because the direction of the inference is from the population to the sample, and not from the sample to the population (Henson & Smith, 2000; Thompson & Snyder, 1997). This is in direct contradiction to how SST is often promoted in textbooks and taught in classrooms. Given that the tests are not about the population, they do not evaluate result replicability in future samples drawn from the population (Thompson, 1996).

Finally, SST presents problems because it restricts the ways in which researchers think about the data. In order to evaluate the practical significance of research, researchers have to be able to make judgments about the results. SST can determine an improbable result, but it cannot determine if an improbable result is an important result (Shaver, 1985). There is little room for researcher judgment; the null is simply accepted or rejected based on the *p* value. Decisions about the importance of the findings are, supposedly, therefore completely objective. Although viewed by many as an advantage

of SST, Thompson (1996) characterized this aseptic treatment of the data as an "escape from the existential human responsibility for making value judgments" (p. 28). Cohen (1994) declared that no matter how convenient a dichotomous accept-reject decision might be, it "is not the way any real science is done" (p. 999). A reject/do not reject strategy turns what ought to be a continuum of uncertainty into a simple dichotomous decision (Kirk, 1996). An accept/do not accept decision is not made by interpretation of or deliberation on the meaning of the data. Instead, the focus is on rejection of the null hypothesis and the *p* value. Kirk (1996) advocated for a broader approach, with the emphasis on the whether the data support the scientific hypothesis. Kirk surmised that physics would not have progressed very far if researchers had merely focused on that fact that *A* was different from *B*. Likewise, Cohen (1994) decried the dependence on *p* values as inhibiting to the development of psychology as a quantitative science.

Voices of Reform

After decades of criticism (Finch et al., 2001), sole dependence on SST is now showing signs of dissipation in light of other data interpretation alternatives. The fourth edition of the *Publication Manual of the American Psychological Association* (1994) noted that "Neither of the two types of probability values reflects the importance or magnitude of an effect because both are dependent on sample size . . . You are encouraged to provide effect size information" (p. 18). However, empirical studies document that this "encouragement" has had little effect on effect size reporting (Henson & Smith, 2000). Following the publication of the fourth edition of the *APA Publication Manual*, the APA convened the Task Force on Statistical Inference (TFSI) in order to evaluate statistical practices and to propose alternatives. With the aim of initiating

discussion on the proposed reforms, the TFSI published a final report in *American Psychologist* (Wilkinson & APA TFSI, 1999). The task force took a firm stand, compelling researchers to "always present effect sizes for primary outcomes" (p. 599, emphasis added). Researchers were encouraged to "add brief comments that place these effect sizes in a practical and theoretical context" (p. 599). The recent fifth edition of the *APA Publication Manual* (2001) reflects the growing reformist agenda. "You are encouraged" (APA, 1994) is now "almost always necessary" (APA, 2001) in regards to reporting effect sizes. The position of the 2001 edition is that "it is almost always necessary to include some index of effect size or strength of relationship in your Results section" (p. 25) and that failure to report effect sizes is a "defect" in research reporting (p. 5). The 2001 edition does not require the reporting of effect sizes, but states that the "general principle to be followed" is to provide readers "with enough information to assess the magnitude of the observed effect or relationship" (p. 26).

Effect Sizes- What Are They?

Without necessarily implying causality, Cohen (1988) explained effect size as "the degree to which the phenomenon is present in the population" (p. 9). Effect sizes give an estimate of the magnitude of differences or relationships (Gall, Borg, & Gall, 1996, p. 188). Important to the purposes of this paper is the clarity given to the practical significance, or usefulness of research from an examination of effect size (Snyder & Lawson, 1993).

Types of Effect Sizes

Effect sizes fall into one of two broad categories (Snyder & Lawson, 1993). The first broad category includes standardized mean difference effects. The statistics in this

category directly examine differences between means. Cohen's *d*, often used in meta-analysis, is an example of a standardized mean effect. Cohen's *d* was the first effect size statistic to be explicitly labeled as such (Kirk, 1996). Based on observed effects in various fields, Cohen (1988) developed operational definitions for small, medium, and large observed effects (.2, .5, and .8, respectively, although these are no more than broad rules-of-thumb and are too rigidly followed). Cohen's unique contribution was making effect sizes useful by providing the researcher with guidelines for interpreting the magnitude of an effect (Kirk, 1996).

The second category of effect size statistics includes those that assess variance-accounted-for effects. Statistics in this category tell how much of the variability in the dependent variable(s) is associated with the variation in the independent variable(s). Examples of variance accounted for statistics are R^2, partial R^2, eta squared, partial eta squared, and omega squared (Snyder & Lawson, 1993). Researchers seeking effects want to be able to account for the largest proportion of variance and/or examine the amount of variance a particular variable explains of the dependent variable. Effect sizes can tell us how much of the dependent variable can be controlled, predicted, or explained by the independent variable(s) (see e.g., Kirk, 1996; Snyder & Lawson, 1993).

Effect Size Interpretation

Routine reporting of effect sizes is a critical first step in assessing the usefulness of research. However, reporting effect size is not the same as interpreting effect size. Most statistical packages print out effect size statistics automatically for analyses such as regression, so acquiring the statistic in these cases is not a problem. Effect sizes for other analyses, such as ANOVA, are not as routinely reported by statistical software packages.

Finch et al. (2001), in an comprehensive review of articles in the *Journal of Applied Psychology* and the *British Journal of Psychology*, found that only a very few reports of effect sizes were given "a substantive interpretation" (p. 202).

Interpretation of effect sizes requires the judgment of the researcher. Snyder and Lawson (1993) illustrated this point with a scenario in which a particular instructional method produced a 5 point increase in the experimental group over the control group on a dependent measure. With a large enough sample size, the results were statistically significant. However, the 5 point difference may or may not be meaningful in the instructional context of the research. With an effect size, the researcher can judge the importance of that difference. Interpretation of effect sizes allows us to focus on the data in the context of the phenomenon being studied (Kirk, 2001).

Reporting and interpreting effect sizes is important in establishing a global context for the research findings at hand: "We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. It enables readers to evaluate the stability of results across samples, designs, and analyses" (Wilkinson & APA TFSI, 1999, p. 599). Reporting effect sizes has implications for future research. Researchers planning follow-up studies need to know the reasonably expected size of an effect in order to plan an appropriate sample size (Hyde, 2001). Reporting effect sizes is also important for meta-analysis of research studies (Wilkinson & APA TFSI, 1999).

Confidence Intervals

In addition to effect sizes, confidence intervals have been advocated as essential in research reporting and interpretation. A confidence interval (CI) is "an interval or

range of plausible values for some quantity or population parameter of interest" (Cumming & Finch, 2001, p. 533). CIs are frequently used and interpreted in a general context: margins of error in an election, range of possible dates for an event in history, and meteorological forecasts ("we expect 8 to 10 inches of snow"). CIs are, as defined by Cumming and Finch (2001), "a best point estimate of the population parameter of interest and an interval about that to reflect likely error--the precision of the estimate" (p. 533). Descriptions of CIs include a statement about the confidence level, usually given as a percentage. The larger the percentage, the greater our confidence that we have captured the population value. It is never certain that the interval has captured the population value, unless the confidence level is set at 100%. Therefore, the width of the interval is an indicator of the precision of the estimate. The price we pay for greater confidence is less precision (Smithson, 2000).

The current APA *Manual* (2001) urges the use of CIs in the interpretation of results:

> The reporting of confidence intervals (for estimates of parameters, for functions of parameters such as differences in means, and for effect sizes) can be an extremely effective way of reporting results. Because confidence intervals combine information on location and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy. (p. 22)

CIs give point and interval information that is accessible and comprehensible (Cumming & Finch, 2001). As the previous examples of interval reporting in weather forecasting, elections, and historical timelines illustrate, a point estimate and a CI use the same unit of measurement as the data (Kirk, 2001). When a familiar scale is used (for example, an

I.Q. scale), a point estimate of a difference and a CI can be used to determine whether findings are trivial, useful, or important (Kirk, 1996). CIs thus promote substantive, reflective, and interpretive thinking about the data on the part of the researcher. Kirk noted that researchers have an obligation to make this kind of judgment: "No one is in a better position than the researcher who collected and analyzed the data to decide whether or not the results are trivial" (p. 755).

CIs provide all the information contained in statistical significance testing, and more (Kirk, 1996, 2001). Cumming and Finch (2001) observed the link between CIs and SST: "Noting that an interval excludes a value is equivalent to rejecting a hypothesis that asserts that value as true" (p. 534). With CIs, the interval is centered correctly on the observed value, rather than on the hypothetical value of the null hypothesis (Hunter & Schmidt, 1990). In addition, CIs provide a range of possible values within which the true difference lies, instead of a dichotomous decision of accept/do not accept (Finch et al., 2001; Kirk, 1996). Some have argued that there is no essential difference between SST and CIs. However, one simple point illustrates that CIs do provide additional information: A CI can be constructed without a null hypothesis.

Confidence Intervals Around Effect Sizes

If effect sizes are advocated for research interpretation, and CIs are also viewed as valuable, then CIs around effect sizes should prove quite useful (Cumming & Finch, 2001; Finch et al., 2001; Wilkinson & APA TFSI, 1999). CIs around effect sizes give us a range of plausible values of the true effect in the population. They also give us a measure of the precision of the estimate. When CIs are constructed around effect sizes and then compared to previous related studies, attention is focused on stability across

studies (Schmidt, 1996). This use of CIs supports meta-analysis and meta-analytic thinking, insufficiently used tools that hold tremendous promise in the social sciences (Cumming & Finch, 2001).

Meta-analysis and Research Synthesis

In the quest for research reporting that focuses on practical and theoretical significance as opposed to statistical significance, meta-analysis has been called a "bright spot on the contemporary scene" (Cohen, 1994, p. 1000). In meta-analysis, results from different studies are translated to a common metric and are then explored in order to discover relationships between studies and findings. Typically, reviews are conducted of the full population of relevant studies. Study outcomes are then transformed to a common metric, usually an effect size (Bangert-Drowns & Rudner, 1991). Gene Glass, credited with first using the term *meta-analysis*, did not conceive of it as a statistical technique, but rather a philosophy or a perspective (Glass, McGaw, & Smith, 1981). According to Glass et al., the aim of meta-analysis was to take research from the ethereal world of statistics to the land of practicality:

> Meta-analysis is aimed at generalization and practical simplicity. It aims to derive a useful generalization that does not do violence to a more useful contingent or interactive conclusion. The world runs on generalizations and marginal utilities. Therein lie many of the difficulties that scientists and men of practical affairs encounter when they meet. (p. 23)

The Power of Meta-analysis vs. the Single Study

Hunter, Schmidt, and Jackson (1982) wrote, "Scientists have known for centuries that a single study will not resolve a major issue . . . . The foundation of science is the

cumulation of knowledge from the results of many studies" (p. 11). In the pursuit of accumulated knowledge, and more specifically, knowledge that is useable, explanatory, and generalizable, meta-analysis holds a distinct advantage over a single study. Cook et al. (1992) summarized the limitations of single studies: (a) generalizable knowledge is rarely gained from a single effort at data collection from a limited sample group, (b) single studies are usually unable to implement more than one variant of a treatment, (c) single studies usually take place at a single time, and (d) available resources limit the breadth of measurement. Glass et al. (1981) illustrated the limited predictability at the individual or study level by noting that actuaries cannot accurately predict if a man will die in the coming year, but they can predict how many persons in a group of 10,000 will die. Wilkinson and the APA TFSI (1999) admonished researchers to practice restraint and consider research results in context: "The thinking presented in a single study may turn the movement of the literature, but the results in a single study are important primarily as one contribution to a mosaic of study effects" (p. 602).

Although a single study may be descriptive of the particular relationship studied, explanation is a more difficult task than description (Cook et. al., 1992). Meta-analysis therefore treats findings from relevant studies as complex data points to be analyzed and understood in a generalizable way (Glass, 1977; Schmidt, 1992). As Schmidt (1992) noted, "Data come to us encrypted, and to understand their meaning we must first break their code" (p. 1179). Meta-analysis serves to clean up and decode research literature, diluting the problematic artifacts of sampling error and measurement error by synthesizing results from many related studies and revealing the cumulative knowledge that is there (Hunter & Schmidt, 1990; Schmidt, 1992).

How, then, does meta-analysis differ from traditional literature review? The misuse of SST has resulted in grave errors in interpretations in review studies. Since studies are considered individually and then compiled, traditional literature reviews often appear to present conflicting results and reviewers falsely conclude that more research is needed (Hunter & Smith, 1990; Schmidt, 1992). Schmidt (1992) illustrated how meta-analysis, using effect sizes from a group of studies, will correctly identify a population value when the majority of the studies found no significance based on SST. Schmidt (1996) further stated that SST actually works against the cumulation of knowledge in such instances.

Meta-analysis and Theory Building

The power of meta-analysis is that it can tell us so much more than an individual study, even more than a group of studies considered individually. Determine the effect of a variable in a sample group, and gain a piece of data on one group. Synthesize effects from many studies, and gain a better understanding of the effect of the variable in the population. When we have solid evidence for the effect in the population, we build theory. Theory in a scientific field is build from consistent, stable research findings. Meta-analysis plays a stabilizing role for theory building because reconceptualization does not have to follow every modest experimental failure (Hedges, 1987). Scientific knowledge grows as we synthesize research findings across time.

Meta-analysis and Replication

If scientific knowledge grows as we synthesize it across time, then scientific knowledge grows more precisely from replication. Scientific knowledge that cannot be replicated is unreliable knowledge (Hubbard & Ryan, 2000; Vacha-Haase, 2001).

Hubbard and Ryan (2000) found that less than 1% of articles in a sample of psychological journals were replications, leading them to suggest that replication is being narrowly defined as exact replication. There are ways to evaluate both external and internal replicability (Thompson, 1996), but meta-analysis provides an additional means of establishing the authenticity of results. Thompson and Snyder (1997) and the APA TFSI (Wilkinson & APA TFSI, 1999) recommended explicitly and reflectively linking research results in a current study to effect sizes in previous studies in order to evaluate result replicability. Likewise, Vacha-Haase (2001) observed that the next best thing to a true replication was "thoughtfully comparing effect sizes in a given study with effect sizes reported in relevant previous literature" (p. 220).

Meta-Analytic Thinking

This paper advocates an approach to research that focuses on the practical significance and usefulness of findings. Interpretation of effect sizes and confidence intervals as well as meta-analytic synthesis of relevant research have been recommended in order to assess the importance and usefulness of a study. The common denominator in these recommended practices is a way of thinking about data: thoughtful, reflective, and global. Meta-analytic thinking focuses on the cumulation of knowledge through the synthesis of prior research and current research with prior research. Through the synthesis of research, we have a much better idea about the population in regard to the phenomenon of interest. Meta-analytic thinking changes the way the individual empirical study is viewed: "It is a new way of thinking about the *meaning* [italics added] of data" (Schmidt, 1992, p. 1173).

Finding Meaning

Cumming and Finch (2001) presented a case for meta-analytic thinking in their *Primer* on CIs. First, meta-analytic thinking necessitates an "accurate and justifiable appreciation of previous research" (p. 555). The TFSI (Wilkinson & APA TFSI, 1999) exhorted researchers to seriously consider prior research and theory. Thompson (1998) decried the failure of researchers to thoughtfully extrapolate expected results from previous literature or theory. Without such thought, science then becomes "an automated, blind search for mindless tabular asterisks using thoughtless hypotheses" (Thompson, 1998, p. 799). This is the antithesis of meta-analytic thinking. Many discoveries and consequent advances in cumulated knowledge have come not from individual studies, but from those who use meta-analytic thinking to find latent meaning in previous research (Schmidt, 1992).

Meta-analytic thinking encourages researchers to report results in a way that makes it easy for other researchers to integrate them into future meta-analyses (Cumming & Finch, 2001). Scientists want to build theory; scientists want to test theory. In order to construct theory, the component parts (relationships between variables) must be synthesized and assembled (Schmidt, 1992). Reporting effect sizes and CIs translates research findings into useable components. When our results are reported in ways that facilitate their use by others, knowledge grows and theory is substantiated. Meta-analytic thinking can be used even on a small scale (Cumming & Finch, 2001). Results from a single researcher, a single laboratory, or a group of collaborators may be integrated using meta-analysis. These small-scale findings can be useful in a larger meta-analysis later on.

Meta-analytic Thinking and Confidence Intervals

Construction of a CI around the effect size of a study gives us a range of plausible values for the true effect in the population and an estimate of the precision of that effect. (Note: Most CIs are constructed using central distributions. However, CIs around effect sizes require use of non-central distributions. See Cumming & Finch [2001] for a more complete discussion.) If we go a step farther and *synthesize* CIs from the effects of a group of related studies, we are thinking meta-analytically about the accumulated data. As a result, we would not only have a global effect, but a global confidence level as well. In other words, synthesizing CIs from related studies gives us the best estimate of the population effect, and the best indication of the precision of that effect.

ESCI and Support for Meta-analytic Thinking

Exploratory Software for Confidence Intervals (ESCI) is a set of interactive simulations that run under Microsoft Excel. ESCI was developed by Cumming (2001) in order to assist researchers in implementing the statistical reforms concerning effect sizes and CIs. Within six workbooks, ESCI demonstrates several key concepts of CIs, including power and repeated sampling. The ESCI software supports the use of standardized effect measures by enabling the researcher to calculate Cohen's *d* for data. Relevant to this discussion is ESCI's support of meta-analysis and meta-analytic thinking using the ESCI workbook MAthinking.

The ESCI software provides graphical displays along with analysis. "Modern statistical graphics" were urged by the TFSI (Wilkinson & APA TFSI, 1999) as a means of encouraging quality conclusions. Graphics, especially with meta-analysis, help the

consumer of research grasp a big picture view as the eye is prompted to compare different pieces of information (Light, Singer, & Willett, 1994).

Using MAthinking

The most basic meta-analysis consists of simply pooling results over studies, weighted by sample size. Using MAthinking, CIs around effect sizes from previous studies are pooled and a CI based on the pooled studies is given. Next, a CI for a current study is found. Then, the CIs from the past studies are pooled with the current study for a final analysis, resulting in the best point estimate of the population effect and level of precision of that effect.

MAthinking uses both original units and standardized (Cohen's *d*) units (Cumming, 2001). In the event that all of the studies we wish to synthesize use the same measurement scale, the original units feature of MAthinking would be appropriate. When different measurement scales are used, data must be converted to a standardized unit of effect size, in this case, Cohen's *d* (Cumming & Finch, 2001). Cohen's *d* enables us to synthesize results from studies that use different sample sizes and different scales, as long as they are measuring the same or very similar variables (Smithson, 2000). Recall that Cohen's *d* is a mean difference statistic (where a difference value is divided by the pooled standard deviations of two groups, much like a *z* score). When there is no difference between groups and the means are the same, *d* equals 0. If there is a difference in group means (an effect), then $d \neq 0$. As *d* becomes more unlike 0, the effect gets larger (assuming in expected directions), indicating larger mean differences relative to the standard deviations.

Demonstration of ESCI

A hypothetical research scenario and a heuristic data set were created in order to demonstrate how MAthinking synthesizes CIs around effect sizes from a group of related studies. Achievement in seventh grade language arts was given as the dependent variable, with hours of participation in a summer enrichment program as the single predictor. Seven past studies and a current study measured language arts achievement using either an overall language score on a standardized achievement test or grade point average in English and reading. Summer enrichment programs also varied, but the overall focus of each program was determined to be similar. The studies varied in number of participants.

The studies were regression studies, with effect size for each reported as R^2 (see Table 1). Based on the assumption that all analyses are correlational in nature, Cohen's *d* (a mean difference effect) can be transformed to into an equivalent Pearson's *r* (or *R*) coefficient, which when squared (R^2), gives us the amount of variance accounted for between the variables. Using Cohen's (1988, p. 23) formula, $d = [2\ (r)]/\ [(1 - r^2)^{.5}]$, a *d* was calculated for each study. Table 1 reports the conversions.

Figure 1 shows a display from MAthinking featuring the research scenario data. The *d* and *n* values for each past study were entered. A CI for each study was calculated when the "set CI" button was triggered. The CI is also pictured graphically as bars around the *d* value. Notice that four of the past research studies captured 0 (not statistically significant), while three others did not capture 0 (statistically significant). Using a traditional manner of examining the literature, one might conclude mixed results concerning the treatment, with the edge going to no effect in the population.

MAthinking next pooled the *d* and *n* values and the CIs from the past research and graphically represented the pooled CI. The pooled effect and CI represents a much more stable estimate of the population effect. The values for the current study were then entered, and a CI was calculated and graphically presented. Note that the current study encompassed the 0, indicating no statistically significant effect. Last, the values and CIs from the past research were pooled with the values and CI from the current study. Although more than half of the past research as well as the current study indicated no statistically significant effect, meta-analytic synthesis of effect sizes (and CI's around the effects) resulted in a different conclusion. This is the power of meta-analytic thinking: conflicted literature may yield a reasonable effect size, with remarkable precision, when examined meta-analytically.

MAthinking graphically reveals that the precision of the estimate of an effect also grows from meta-analytic treatment of data. Recall that the width of the interval is an indicator of the precision of the estimate (wider intervals, less precision; narrower intervals, more precision). Using the MAthinking graphics (Figure 1), compare the individual widths of the CIs from the past research to the CI width for the past research-pooled. The past research-pooled CI is less than half of the width of the any one of the individual past studies. Likewise, the CI width for the present study is more than twice the width of the CI width for past research-pooled. Notice the impact on CI width after pooling the current study with past research. Additional, though modest, shrinkage of the CI width occurred again after pooling of past and present research studies. ESCI and the graphics of MAthinking support the meta-analytic thinking that allows us to hone the

precision of an effect estimate. Instead of a singular effect from an isolated study, we can have a global effect and a global confidence level.

More Insights Using ESCI

Using MAthinking, we can discover additional insights from research studies (Cumming & Finch, 2001). First, a current study will have only a modest influence on the final picture in a meta-analysis unless the *n* is particularly large. Second, the most influential component in this analysis is the magnitude of the effect size, and especially the consistency of the effect sizes over the studies. Bigger effect sizes and larger sample sizes decrease the possibility that the interval will capture 0. MAthinking also gives insight into the statistical significance status of individual studies. The MAthinking worksheet allows the display to show statistical significance test results (see Figure 2). Several studies, including the current study, did not reach statistical significance. Yet when studies were combined meta-analytically, statistical significance was reached. This evidence provides justification for the, at present, rare practice of reporting non-statistically significant results in journals.

Conclusion

The practical significance, usefulness, and generalizability of research has for decades hinged on a finding of statistical significance. Voices of reform, as evidenced in the in the TFSI report (Wilkinson & APA TFST, 1999) and revisions to the APA *Manual* (2001), have called for consistent use of effect sizes, CIs, and meta-analytic synthesis of literature as a superior means of assessing these three criteria. Researchers are called upon to think meta-analytically: looking at their results in a new way, thinking reflectively, and interpreting them in the global context of past studies. The ESCI

software supports and graphically illustrates such meta-analytic thinking. By synthesizing effect sizes, and the CIs around them, past research findings reveal latent knowledge and theory that might never have been discovered if the studies had been considered individually. Current studies can be synthesized with past studies for a global effect. Through active, reflective interpretation and placement of data in a global context, meta-analytic thinking guides the scientist in search for meaning outside one's own research setting.

Author Note

ESCI can be obtained at www.psy.latrobe.edu.au/esci by rapid download at a minimal cost.

References

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Bailar, J. C., & Mosteller, F. (1988). Guidelines for statistical reporting in articles for medical journals. *Annals of Internal Medicine, 108*, 266-273.

Bangert-Drowns, R. L., & Rudner, L. M. (1991). *Meta-analysis in educational research.* Washington, DC: ERIC Clearinghouse on Tests Measurements and Evaluation. (ERIC Document Reproduction Service No. ED 339 748)

Cohen, J. (1988). *Statistical power analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist, 49(12)*, 997-1003.

Cook, T. D., Cooper, H., Corday, D. S., Hartmann, H., Hedges, L. V., Light, R. J., et al. (1992). *Meta-analysis for explanation.* New York: Russell Sage Foundation.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2*, 161-172.

Cumming, G. (2001). *Exploratory software for confidence intervals* [Computer software and manual]. Retrieved from http://www.psy.latrobe.edu.au/esci/

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61(4)*, 532-574.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the

Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement, 61(2)*, 181-210.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research*. White Plains, NY: Longman.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science: The empirical cumulativeness of research. *American Psychologist, 42*, 443-455.

Henson, R. K., & Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the APA task force report and current trends. *Journal of Research and Development in Education, 33(4)*, 285-296.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology-and its future prospects. *Educational and Psychological Measurement, 60(5)*, 661-681.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting bias in research findings*. Newberry Park: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Culminating research findings across studies*. Beverly Hills, CA: Sage.

Hyde, J. S. (2001). Reporting effect sizes: The roles of editors, textbook authors, and publication manuals. *Educational and Psychological Measurement, 61(2)*, 225-228.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56(5)*, 746-759.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61(2)*, 213-218.

Kover, A. J. (2000). A response to the Hubbard and Ryan article. *Educational and Psychological Measurement, 60(5)*, 691-692.

Light, R. J., Singer, J. D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analysis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 439-452). New York: Russell Sage Foundation.

Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta- analysis, and cumulative knowledge in psychology. *American Psychologist, 47(10)*, 1173-1181.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*(2), 115-129.

Shaver, J. (1985). Chance and nonsense. *Phi Delta Kappan, 67(1),* 57-60.

Smithson, M. (2000). *Statistics with confidence*. Thousand Oaks, CA: Sage.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61(4)*, 334-349.

Stewart, D. W. (2000). Testing statistical significance testing: Some observations of an agnostic. *Educational and Psychological Measurement, 60(5)*, 685-690.

Thompson, B. (1994). *The concept of statistical testing.* Washington, DC: ERIC

Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No. ED 366 654)

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*, 26-30.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53*, 799- 800.

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in *The Journal of Experimental Education. The Journal of Experimental Education, 66(1)*, 75-83.

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement, 61(2)*, 219-224.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Table 1

Effect Sizes of Past, Current, and Pooled Studies

Study	*n*	*R*	R^2	*d*
Study 1	21	.50	.25	1.15
Study 2	9	.28	.08	.58
Study 3	12	.27	.07	.55
Study 4	30	.53	.28	1.25
Study 5	28	.17	.03	.34
Study 6	9	.20	.04	.40
Study 7	22	.33	.11	.71
Past Studies, Pooled	131	.36	.13	.78
Current Study	16	.22	.05	.45
Past + Current, Pooled	147	.35	.12	.74

Figure Captions

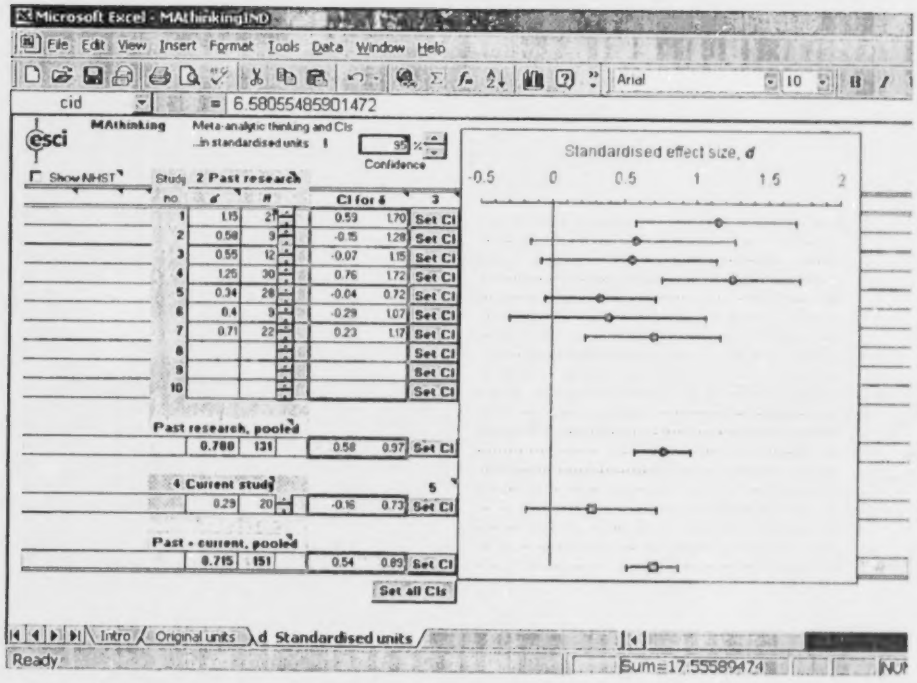
Figure 1. A partial image from the standardized units sheet of MAthinking.

Note. The *d* and *n* values for the past studies were entered at the top left. Clicking on the gray buttons ("set CI") prompted ESCI to calculate the CI for each *d*. The results were given as a CI and pictured graphically as bars around the *d* value. A similar analysis is shown for the past studies, pooled. A *d* and *n* for the current study was entered, and a CI was calculated and graphed. Finally, an analysis is shown for the past studies + the current study.

Figure 2. A partial image from the standardized units sheet of MAthinking.

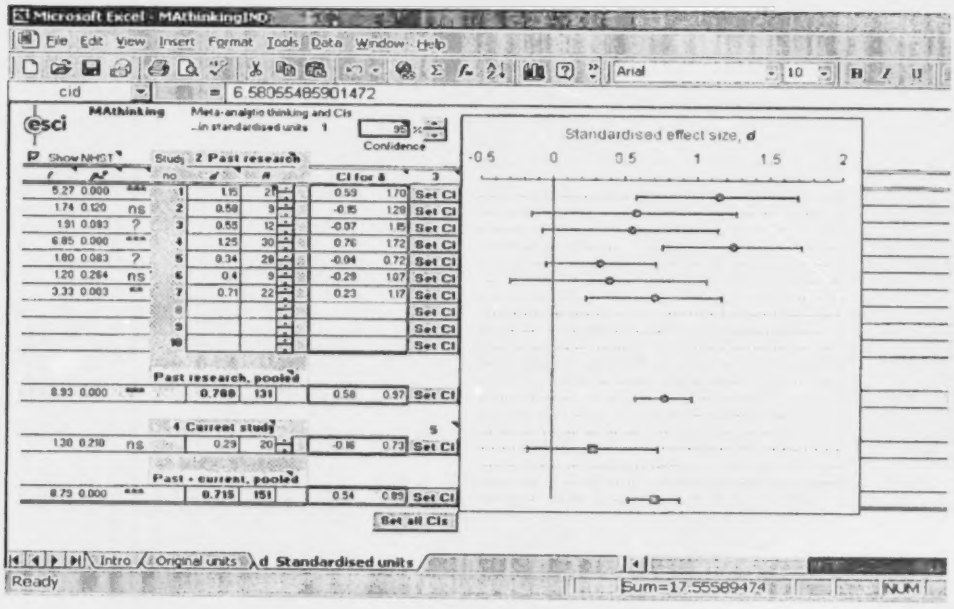
Note. By clicking on the "show NHST" button (top left corner), MAthinking calculated the results of a statistical significance test for each study and the pooled studies.

Figure 1



BEST COPY AVAILABLE

Figure 2



BEST COPY AVAILABLE

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

TM034287

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Meta-analytic Thinking: Using ESCI to Synthesize Effects Across Studies

Author(s): Janet Ray

Corporate Source: University of North Texas

Publication Date: February 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Janet Ray

Printed Name/Position/Title: Janet Ray / Research Associate

Organization/Address: 3208 Hunter Lane, Plano, TX 75093

Telephone: 940-369-8385

FAX: 940-565-2185

E-Mail Address:

Date: 1/28/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com